EM Algorithms for Extreme-Value Software Reliability Models

Hiroyuki Okamura
Department of Information Engineering,
Graduate School of Engineering,
Hiroshima University,
1–4–1 Kagamiyama,
Higashi-Hiroshima 739–8527, Japan
okamu@rel.hiroshima-u.ac.jp

Tadashi Dohi
Department of Information Engineering,
Graduate School of Engineering,
Hiroshima University,
1–4–1 Kagamiyama,
Higashi-Hiroshima 739–8527, Japan
dohi@rel.hiroshima-u.ac.jp

Abstract

This paper considers the software reliability model where each fault-detection time is distributed in accordance with the extreme-value distribution. We first introduce the basic software reliability model with doubly exponential (Gumbel) distribution, and show that the other extreme-value software reliability models can be derived by changing the sample of fault-detection data. Furthermore, we provide the efficient methods to compute the maximum likelihood estimates of the model parameters by using the EM (expectation-maximization) algorithm.

1 Introduction

During the last three decades, the software reliability models (SRMs) have been developed by many authors, and some of them have been used in the actual software testing phase. Among the SRMs, the non-homogeneous Poisson process (NHPP) models play a central role to assess the software reliability, the number of remaining faults in the software, the optimal software release schedule, *etc.*

This paper considers the software reliability model where the fault-detection time is distributed in accordance with the extreme-value distribution. We first introduce the basic software reliability model with doubly exponential (Gumbel) distribution, and show that the other extreme-value software reliability models can be derived by changing the samples of fault-detection data. For the parameter estimation problem, we develop the EM (expectation-maximization) algorithm for the SRM based on the Gumbel distribution, and show that this algorithm can be applied to the other extreme-value SRMs.

2 Extreme-Value Software Reliability Models

Let $N(t)$ denote the number of software faults detected by time t. An NHPP-based SRM makes the following assumptions:

Assumption A: Software failures caused by software faults occur at independent and identically distributed (i.i.d.) random times having the continuous probability distribution function $F(t)$ with density $f(t) = dF(t)/dt$.

Assumption B: The initial number of software faults, N (> 0), is finite.

Under the above assumptions, the probability mass function of the number of faults detected by time t is given by

$$\Pr\{N(t) = n|N\} = \binom{N}{n} F(t)^n \overline{F}(t)^{N-n}, \tag{1}$$

where $\overline{F}(\cdot) = 1 - F(\cdot)$. If the number of initial fault contents N is unknown, it will be appropriate to consider that N is a discrete (integer-valued) random variable. Langberg and Singpurwalla (1985) prove that when the initial number of software faults N obeys the Poisson distribution with parameter ω (> 0), the number of software faults experienced before time t is given by the following NHPP:

$$\Pr\{N(t) = k\} = \frac{\{\omega F(t)\}^k}{k!} e^{-\omega F(t)}. \tag{2}$$

Equation (2) is equivalent to the probability mass function of the NHPP having the mean value function $\omega F(t)$. From this modeling framework, most NHPP-based SRMs can be derived by choosing the software fault-detection time distribution $F(t)$. If $F(t) = 1 - \exp\{-\beta t\}$ ($\beta > 0$), then we can derive the Goel and Okumoto model (Goel and Okumoto 1979) with mean value function $\mathrm{E}[N(t)] = \Lambda(t) = \omega(1 - \exp\{-\beta t\})$.

In this paper, we try to apply the extreme-value distribution into the software fault-detection time distribution. In general, there are some types of the extreme-value distribution. Since the extreme-value distribution for minimum can be reduced to the extreme-value distribution for maximum, we first treat the extreme-value distribution for maximum, namely, the Gumbel distribution (Type I extreme-value distribution). The Gumbel distribution function is given by

$$F(t; \mu, \theta) = \exp\left\{-\exp\left[-\left(\frac{t-\mu}{\theta}\right)\right]\right\}, \tag{3}$$

where μ (> 0) and θ (> 0). However, the domain of the Gumbel distribution is $t \in (-\infty, \infty)$ and therefore it cannot be directly used as the software fault-detection time distribution. To change the domain of the Gumbel distribution, we give two approaches: truncation approach and logarithm approach. First, by truncating the Gumbel distribution at the origin, we have

$$F_{trunc}(t; a, b) = \frac{a^{b^t} - a}{1 - a}, \tag{4}$$

where $a = \exp\{-\exp(\mu/\theta)\}$ and $b = \exp\{-1/\theta\}$. The curve of the distribution function draws the Gompertz curve. Next, we introduce the logarithm approach. Let X be the random variable which obeys the Gumbel distribution. Define $Y = e^X$. Then the random variable Y can take positive values and obeys the following distribution function:

$$F_{log}(t; \alpha, \beta) = \exp\left\{-\beta t^{-\alpha}\right\}, \tag{5}$$

where $\alpha = 1/\theta$ and $\beta = \exp(\mu/\theta)$. This form is equivalent to the Fréchet distribution with support $t > 0$.

On the other hand, the extreme-value distribution for minimum can be derived by letting $Y = -X$, which is given by

$$G(t; \mu, \theta) = 1 - \exp\left\{-\exp\left(\frac{t+\mu}{\theta}\right)\right\}. \tag{6}$$

Similarly, the truncation and logarithm approaches can be applied to the extreme-value distribution for minimum. By the truncation approach, we obtain

$$G_{trunc}(t; \alpha, \lambda) = 1 - \exp\left\{\frac{\lambda}{\alpha}[1 - \exp(\alpha t)]\right\}, \tag{7}$$

where $\lambda = \exp(\mu/\theta)/\theta$. Also, the logarithm approach yields

$$G_{log}(t; \alpha, \beta) = 1 - \exp\left\{-\beta t^{\alpha}\right\}. \tag{8}$$

These distribution functions in Eqs. (7) and (8) distribution functions also correspond to the Gompertz distribution and the Weibull distribution, respectively.

Consequently, substituting the truncated or logarithmic distribution into $F(t)$, we have the four types of extreme-value SRM. Notice that, since the logarithmic extreme-value distribution for minimum is the Weibull distribution, the corresponding SRM is reduced to the generalized exponential SRM by Goel (1985).

3 Parameter Estimation

The maximum likelihood estimates (MLEs) are given by the parameters which maximize the log-likelihood function (LLF) for provided data. Thus, in the maximum likelihood estimation, we find the parameters which satisfy the first-order condition of optimality for the LLF, namely the simultaneous likelihood equations.

Since the likelihood equations are non-linear equations, any iterative algorithm such as the Newton's method is used to calculate the parameters satisfying the likelihood equations. In estimating the parameters, we always take care of constraints of model parameters. The model parameters in the SRMs are usually subject to an implicit constraint such as positive condition. However, it should be noted that the Newton's method and the other numerical methods do not always converge to MLEs satisfying the constraints, if the initial values in the algorithms are far from the MLEs. This property is called the local convergence. This property causes the difficulty on the choice of initial values in the parameter estimation. To overcome the problem on selecting initial values, Okamura *et al.* (2002, 2003) introduce the EM algorithms for the SRMs which are modeled in the framework mentioned in the previous section. In this paper, we apply the EM algorithm to the SRM with the Gumbel distribution.

Suppose that the time domain data on the software fault-detection, $s_1, \ldots, s_k, t_{obs}$ are available, where s_i and t_{obs} denote the software fault-detection time and the observation time, respectively. In this paper, we develop the EM algorithm which is applied to four kinds of extreme-value SRM. Consider the modified fault data, $\mathcal{D}_{max} = (s_1, \ldots, s_k, t_{obs})$, $\mathcal{D}_{min} = (-s_1, \ldots, -s_k, -t_{obs})$, $\mathcal{D}_{log,max} = (\log s_1, \ldots, \log s_k, \log t_{obs})$ and $\mathcal{D}_{log,min} = (-\log s_1, \ldots, -\log s_k, -\log t_{obs})$. Assuming that the changed data are sampled from the Gumbel distribution, the original data $s_1, \ldots, s_k, t_{obs}$ obey respective types of extreme-value SRMs, so that we can estimate the parameters for four types of extreme-value SRM by developing the EM algorithm with the Gumbel distribution.

The EM algorithm is an iterative method for the estimation problem with incomplete data. There are two parts: E-step and M-step. In the E-step, we calculate the expected value of the LLF for complete data, provided that incomplete data is observed. Calculating the expected LLF requires the model parameters, but provisional parameters are used as the model parameters in most cases. In the M-step, we find the parameters so as to maximize the expected LLF calculated in the E-step. After finding the parameters which maximize the expected LLF, the provisional parameters are renewed by the parameters. By executing the E-step and the M-step iteratively until the provisional parameters converge to certain points, we get the MLEs for model parameters.

In this case, all the fault data $\mathcal{D}_{max}$, $\mathcal{D}_{min}$, $\mathcal{D}_{log,max}$ and $\mathcal{D}_{log,min}$ are obviously incomplete data because all of them are truncated by the time t_{obs}. In particular, $\mathcal{D}_{max}$ and $\mathcal{D}_{min}$ are also truncated by the time 0. Therefore, we derive the expected LLF as follows.

Suppose that $\mathcal{D} := \mathcal{D}_{max}$, $\mathcal{D}_{min}$, $\mathcal{D}_{log,max}$ or $\mathcal{D}_{log,min}$. Then we have

$$\mathrm{LLF}(\omega, \mu, \theta|\mathcal{D}) = -\mathrm{E}[N|\mathcal{D}]\log\theta - \mathrm{E}\left[\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\theta}\right)\middle|\mathcal{D}\right] - \mathrm{E}\left[\sum_{i=1}^{N}\exp\left\{-\left(\frac{X_i-\mu}{\theta}\right)\right\}\middle|\mathcal{D}\right] \tag{9}$$

where N is the total number of software faults and X_i, $i = 1, \ldots, N$ are the fault-detection times for all the faults.

The usual EM algorithm requires the closed form solutions of the likelihood equations for the fault-detection time distribution, so that the closed form constructs the update formulae in the M-step. However, in the case of the Gumbel distribution, it is not easy to find the closed form solutions. Thus we apply the generalized EM (GEM) algorithm to estimate the parameters (McLachlan and Krishnan 1997). The GEM algorithm does not always require the closed form solutions of the likelihood equations, but the corresponding update in the M-step has to make the expected LLF increase. Applying the GEM algorithm, we derive the following update formulae (M-step) to estimate the parameters:

$$\omega := \mathrm{E}[N|\mathcal{D};\omega',\mu',\theta'], \tag{10}$$

$$\mu := \theta'\log\left(\frac{\mathrm{E}[N|\mathcal{D};\omega',\mu',\theta']}{\mathrm{E}\left[\sum_{i=1}^{N}\exp\left[-(X_i/\theta')\right]\middle|\mathcal{D};\omega',\mu',\theta'\right]}\right), \tag{11}$$

$$\theta := \theta'\mathrm{E}\left[\sum_{i=1}^{N}\left(\frac{X_i-\mu'}{\theta'}\right)\left\{1-\exp\left[-\left(\frac{X_i-\mu'}{\theta'}\right)\right]\right\}\middle|\mathcal{D};\omega',\mu',\theta'\right]/\mathrm{E}[N|\mathcal{D};\omega',\mu',\theta']. \tag{12}$$

Applying the following formulae (E-step) to the above, we can derive the EM algorithm for four types of extreme-value SRM:

For any function $h(\cdot)$,

$$\mathrm{E}\left[\sum_{i=1}^{N} h(X_i)\middle|\mathcal{D}_{max};\omega',\mu',\theta'\right] = \sum_{i=1}^{n} h(x_i) + \omega'\left(\int_{-\infty}^{0} h(x)f(x)dx + \int_{t_{obs}}^{\infty} h(x)f(x)dx\right), \quad (13)$$

$$\mathrm{E}\left[\sum_{i=1}^{N} h(X_i)\middle|\mathcal{D}_{min};\omega',\mu',\theta'\right] = \sum_{i=1}^{n} h(-x_i) + \omega'\left(\int_{-\infty}^{-t_{obs}} h(x)f(x)dx + \int_{0}^{\infty} h(x)f(x)dx\right), \quad (14)$$

$$\mathrm{E}\left[\sum_{i=1}^{N} h(X_i)\middle|\mathcal{D}_{log,max};\omega',\mu',\theta'\right] = \sum_{i=1}^{n} h(\log x_i) + \omega'\int_{\log t_{obs}}^{\infty} h(x)f(x)dx, \quad (15)$$

$$\mathrm{E}\left[\sum_{i=1}^{N} h(X_i)\middle|\mathcal{D}_{log,min};\omega',\mu',\theta'\right] = \sum_{i=1}^{n} h(-\log x_i) + \omega'\int_{-\infty}^{-\log t_{obs}} h(x)f(x)dx. \quad (16)$$

Note that, in the case of $\mathcal{D}_{max}$ and $\mathcal{D}_{min}$, the parameter ω is finally given by $\omega = \omega' F(0)$ and $\omega = \omega'\overline{F}(0)$, respectively.

4 Conclusion

In this paper, we have discussed four types of extreme-value SRM, which are based on the familiar software debugging theory. Furthermore, we have developed the efficient iterative scheme to calculate the MLEs of the model parameters. The proposed estimation algorithms are based on the EM principle. Although the algorithm is proposed to the specified SRM based on the Gumbel distribution, we can apply the algorithm to the other extreme-value SRMs by changing the observed samples.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture: Grant-in-Aid for Young Scientists (B), Grant No. 15700060 (2003-2004) and Exploratory Research, Grant No. 15651076 (2003-2005).

References

Goel, A. L. (1985). Software reliability models: assumptions, limitations and applicability. *IEEE Trans. Software Eng.*, **SE-11**, 1411–1423.

Goel, A. and Okumoto, K. (1979). Time-dependent error-detection rate model for software reliability and other performance measures. *IEEE Trans. Reliab.*, **R-28**, 206–211.

Langberg, N. and Singpurwalla, N. D. (1985). Unification of some software reliability models. *SIAM J. Sci. Comput.*, **6**, 781–790.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

Okamura, H., Watanabe, Y. and Dohi, T. (2002). Estimating mixed software reliability models based on the EM algorithm. *Proc. Int'l Sympo. on Empirical Software Eng.*, 69–78.

Okamura, H., Watanabe, Y. and Dohi, T. (2003). An iterative scheme for maximum likelihood estimation in software reliability modeling. *Proc. 14th Int'l Sympo. on Software Reliab. Eng.*, 246–256.